EXHIBIT AV

Notes on Contributors

ELIZABETH J. ALLAN is an assistant professor of Higher Educational Leadership at the University of Maine in Orono. Her current research and teaching interests center around campus climates in postsecondary education, as well as student development, classroom teaching practices, and college student experiences outside the classroom. She is the author of "Hazing in High School and College" in *Encyclopedia of Men and Masculinities* (edited by M. Kimmel and A. Aronson, forthcoming), and "Gender and Hazing: Analyzing the Obvious" in *Hazing* (edited by H. Nuwer, 2002).

ROY O. FREEDLE is a former senior research psychologist; he retired in 1998 from the Educational Testing Service in Princeton, New Jersey. His primary professional focus is the application of discourse theory to the study of reading comprehension, essay writing, and the effects of culture on language evaluation. He is editor-in-chief of a 65-volume book series, *Advances in Discourse Processes* (1978–present). His other books include *Language Comprehension and the Acquisition of Knowledge* (with John B. Carroll, 1992) and *Artificial Intelligence and the Future of Testing* (1990).

JOSIANE HUDICOURT-BARNES is a research and development specialist at the Chèche Konnen Center of TERC, a not-for-profit education research and development organization based in Cambridge, Massachusetts. She is interested in bilingual education, science education, and cognitive development. She is coauthor of "Rethinking Diversity in Learning Science: The Logic of Everyday Languages," with B. Warren, C. Ballenger, M. Ogonowski, and A. Rosebery, in the *Journal of Research in Science Teaching* (2001), and "The Sound of Drums," with F. Conant, A. Rosebery, and B. Warren, in *Classroom Diversity: Connecting Curriculum to Students' Lives* (edited by E. McIntyre, 2001).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Correcting the SAT's Ethnic and Social-Class Bias: A Method for Reestimating SAT Scores

ROY O. FREEDLE

The SAT has been shown to be both culturally and statistically biased against African Americans, Hispanic Americans, and Asian Americans. In this article, Roy Freedle argues for a corrective scoring method, the Revised-SAT (R-SAT), to address the nonrandom ethnic test bias patterns found in the SAT. The R-SAT, which scores only the "hard" items on the test, is shown to reduce the mean-score difference between African American and White SAT test-takers by one-third. Further, the R-SAT shows an increase in SAT verbal scores by as much as 200 to 300 points for individual minority test-takers. Freedle also argues that low-income White examinees benefit from the revised score as well. He develops several cognitive and cultural hypotheses to explain the ethnic regularities in responses to various test items. Freedle concludes by offering some predictions as to how ethnic populations are likely to be affected by the new designs currently being proposed for the SAT, and describes the implications of the R-SAT for increasing minority admission to select colleges.[1]

Statement of Problem

In theory, the educational movement that initiated standardized testing for purposes of college admission originally held the promise of identifying students of merit from diverse social-class and ethnic backgrounds who otherwise would not have been considered for admission into the nation's select colleges (Lemann, 1999). But, in practice, this early promise has not been fulfilled, especially for minority groups whose mean test performance has departed significantly from White mainstream test-takers. Over the past several decades, the search for a more equitable ethnic representation in our nation's select colleges led to the adoption of affirmative action, a policy that is increasingly under attack (Lemann, 1999). The recent erosion of affirma-

Harvard Educational Review Vol. 73 No. 1 Spring 2003
Copyright © by the President and Fellows of Harvard College

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

tive action increases the need for other means of identifying promising minority students for admission into our system of higher education. This article suggests one avenue for solving this problem.

The chief purpose of this article is to present a new method of scoring the SAT (called the Revised-SAT, or R-SAT) that will greatly increase the number of high-scoring minority individuals. This new scoring method corrects for two types of potential test bias: cultural and statistical. It has the potential to justify the acceptance of many more minority individuals into select colleges based on their test performance, along with other important factors such as high school grade point average.

Stephen Jay Gould (1995) reminds us that a test can be biased in at least two ways, culturally or statistically. Gould's distinction crystallizes several ideas regarding test bias. He explains that a standardized test may be culturally biased when one group (typically a minority population) performs *consistently* lower than some reference population — typically, the White population. He adds that a test is considered statistically biased if two individuals (e.g., one African American, one White) who get the same test score nevertheless perform differently on some criterion external to the test, such as school grades.

One can extend Gould's argument to say that a test is culturally biased if individuals from different ethnic groups interpret critical terms in many of the test items differently. The consequence of this interpretive difference would be the observed mean differences in test performance. Building as well on Gould's statistical bias definition, there is at least one other sense in which a test can be biased. For example, if two individuals get the same verbal test score, it is reasonable to assume that they should perform approximately equally well on all aspects of the verbal test itself. However, if two individuals with the same overall score — or, more generally, two ethnic groups matched on some total test score — should differ substantially on different subparts of a test, we would say that the test is also statistically biased.

Based on the results presented below, I assert that the SAT, as currently administered and analyzed, is both *culturally* and *statistically* biased in the ways described above. I also show how both cultural and statistical bias can be partially ameliorated by scoring one half of the SAT, the *hard* part. The *hard* items are those that are often dependent on rare vocabulary, whereas the *easy* items are dependent on terms that are typically more common. (I explain this in more detail below.)

After developing a rationale for this new scoring, I argue that this method for reestimating at least the verbal SAT score can have significant implications for increasing minority admission into select colleges that often prefer very high verbal SAT scores.[2] I also show that this new method carries positive import for many White individuals from disadvantaged backgrounds (e.g., low-income individuals and/or individuals whose preferred language is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

not English).[3] Therefore, a revised SAT score (R-SAT) could benefit individuals from various ethnic groups.

Background Studies of Cultural and Statistical Bias in the SAT

Cognitive psychologists Freedle and Kostin (1987, 1988, 1990, 1997) use a measure called the standardized Differential Item Functioning (DIF) statistic to study ethnic bias in any standardized test. Briefly, DIF is a statistical procedure (Dorans & Kulick, 1983, 1986) that examines minority and White responses to each test item (both verbal and quantitative).[4] The first step in conducting the DIF procedure is to look at performance on the first test item for all examinees who scored 200 on the verbal SAT. The respective proportions of White and African Americans who correctly answered that first item are computed, and then the difference in proportions is determined. This difference is then weighted by the number of African Americans scoring 200. Notice that both Black and White students are said to be of matched ability here because both groups received the same score of 200.

Next, for the same item, the same weighted computation is performed for all White and African American candidates who scored a 210. This continues for all score levels, through 800. These steps (from 200 to 800) together yield sixty-one weighted computations, all applying to the first item. One sums these sixty-one computations and determines their average value. This average is called the DIF score for the first item. All subsequent verbal items are then examined by the same procedure, and a DIF score is assigned to each of them. A positive DIF score for an item indicates that the African American population performs differentially better than their matched-ability White peers. A negative DIF score for an item indicates that the African American population performs differentially less well than their matched-ability White counterparts.[5,6]

Freedle and Kostin (1988) show that there is evidence of an unintended but persistent cultural and statistical bias in the verbal section of the SAT that adversely affects African Americans.[7] Specifically, using the DIF method, these studies show that Whites tend to score better on *easy* items and African Americans on *hard* items.[8] It should be noted from the outset that virtually all these DIF item effects are typically small. For example, White students may get 84 percent correct on some *easy* items, while African Americans get a slightly lower number, say 82 percent, correct for that same item. Conversely, for some particular *hard* items, White students might get 30 percent correct whereas African Americans might get a slightly higher score, say 31 percent correct. What is unusual about these effects is their highly patterned nature; that is, many *easy* items show a small but persistent effect of African Americans' underperformance, while many *hard* items show their overperformance.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Later I will make it clear that these small single-item bias effects become magnified, partly because the traditional scoring of a paper-and-pencil SAT gives equal weight to every item. In other words, a correct *easy* item carries the same weight as a *hard* item. I examine this assumption of equal weight below in terms of its effects on ethnic bias.

The Lexical Ambiguity and Cultural Familiarity Hypothesis

The largest positive and negative DIF values occurred among analogies and antonyms, the verbal item types with the least verbal context. In contrast, the smallest DIF values tended to occur for the reading comprehension items, which had the maximum verbal context. Freedle and Kostin (1990) observed two major effects: 1) DIF item values, both positive and negative, increase as the amount of verbal context decreases; and 2) within each item type (except for the reading items) the *easy* items typically receive negative DIF values, while the *hard* items typically receive positive DIF values. In short, test bias against minorities occurs primarily for *easy* analogy and *easy* antonym items.

Two alternative ways to report on the relationship between item difficulty and DIF are displayed in Table 1. First, we sometimes resort to a shorthand way of reporting the contrast between *hard* and *easy* items by presenting just two values: one for the grouped *easy* items (in this case –.027) and a second for the grouped *hard* items (in this case +.012). These two values typically establish the pattern that most of the *easy* items yield a negative DIF (favoring the Whites), while the *hard* items yield a positive DIF (favoring the African American examinees).[9]

Second, a shorthand way to report the relationship between item difficulty and DIF is to compute a single correlation between item difficulty (the values that range from 4.0 through 18.9 in Table 1) and DIF. As shown in Table 1, the correlation for analogies is equal to .52 for all 217 analogies. In this case, the fact that the correlation is positive indicates that *hard* items are associated with positive DIF values, while *easy* items are associated with negative DIF values.

Replication of the Ethnic Bias Pattern: A Brief Review

Freedle and Kostin (1997) reviewed other background studies that replicated and extended the general pattern of DIF that they reported earlier. Kulick and Hu (1989) later reported positive DIF scores favoring African Americans, Hispanic Americans, and Asian Americans on *hard* item performance, not only for verbal SAT items, but also for the quantitative SAT-M items.[10,11,12] Kulick and Hu also replicated an overall positive DIF result for the *hard* verbal items. They also found, like Freedle and Kostin, that analogies and antonyms yielded the largest DIF scores for each ethnic group, and reading items yielded the least bias. Schmitt, Dorans, Crone, and Maneck-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 1 *Average DIF Values for Different Levels of Item Difficulty for Three SAT Verbal Item Types and the Correlation between DIF and Item Difficulty*

Item Difficulty	*Analogies*	*Antonyms*	*Sentence Completions*	*Reading Comprehension*
(Easiest)[++]				
4.0–4.9	–.016 (1)[+]	–.016 (4)	.0 (0)	.0 (0)
5.0–5.9	–.033 (9)	–.026 (10)	–.015 (8)	.0 (0)
6.0–6.9	–.035 (8)	–.018 (17)	–.044 (10)	–.018 (5)
7.0–7.9	–.040 (16)	–.011 (23)	–.014 (17)	.003 (11)
8.0–8.9	–.029 (26)	–.017 (17)	.003 (20)	–.001 (19)
9.0–9.9	–.024 (17)	–.024 (26)	.003 (17)	.008 (23)
10.0–10.9	–.010 (21)	.023 (21)	–.015 (12)	.008 (46)
Average for *Easy*				
4.0 – 10.9	–.027 (98)	–.011 (118)	–.010 (84)	.005 (104)
(Hardest)				
11.0–11.9	.004 (13)	.012 (15)	.003 (7)	.009 (47)
12.0–12.9	–.006 (16)	.018 (17)	.010 (11)	.004 (48)
13.0–13.9	.018 (27)	.017 (32)	.014 (16)	.008 (33)
14.0–14.9	.013 (40)	.016 (43)	.010 (20)	.005 (20)
15.0–15.9	.017 (12)	.023 (32)	.015 (19)	.012 (19)
16.0–18.9	.019 (11)	.026 (14)	.012 (5)	.014 (4)
Average for *Hard*				
11.0–18.9	.012 (119)	.018 (153)	.012 (78)	.007 (171)
Correlation between DIF and all Difficulty Levels:	.52 ***	.41***	.48***	.08 ns

*** Significant, p < .001 (2–tailed); ns = not significant (p > .05)

[+] The number of items used to compute each average DIF value is in parentheses.

[++] Item Difficulty values are actually Equated Delta values. These later values reflect a statistical procedure that standardizes the item difficulties across individual test forms in such a way that difficulty is expressed on a common scale, thereby allowing for items to be meaningfully pooled across different test forms (for further details, see Kulick & Hu, 1989, p. 3).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

shana (1991) also reported similar verbal DIF findings for all three ethnic groups. Finally, Freedle and Kostin (1997) reanalyzed data reported by Raju, Drasgow, and Slinde (1993) regarding a 45-item vocabulary test for grades ten and twelve. Participants included 245 African American and 436 White students. Again, they found significant effects favoring the African American students for the *hard* vocabulary items and disfavoring the African Americans for the *easy* items.

It is clear then that other researchers have replicated this ethnically based response pattern, with *easy* items generally being better performed on by the White majority and *hard* items generally being better performed on by each of the minority groups. What is needed now is a more explicit cognitive interpretation that helps to explain this highly replicable test bias pattern.

The Cultural Unfamiliarity Hypothesis

A culturally based interpretation helps explain why African American examinees (and other minorities) often do better on many *hard* verbal items but do worse than matched-ability Whites on many *easy* items. To begin with, *easy* analogy items tend to contain high-frequency vocabulary words while *hard* analogy items tend to contain low-frequency vocabulary words (Freedle & Kostin, 1997). For example, words such as "horse," "snake," "canoe," and "golf" have appeared in several *easy* analogy items. These are words used frequently in everyday conversations. By contrast, words such as "vehemence," "anathema," "sycophant," and "intractable" are words that have appeared in *hard* analogy items, and do not appear in everyday conversation (Berger, 1977). However, they are likely to occur in school-related contexts or in textbooks. In fact, these rare words do occur in a large sample of college textbooks analyzed by Breland and Jenkins (1997).

Let's consider some possible psychological dimensions associated with word frequency. It is well known that common words often have many more semantic (dictionary) senses than rare words. For example, Freedle and Kostin (1990) report that fifteen high-frequency analogy words (such as "horse" and "snake") had an average of 5.2 dictionary entries, whereas rare analogy words (such as "vehemence" and "anathema") had an average of only 2.0 dictionary entries. Various researchers have hypothesized that each cultural group assigns its own meanings to such common words to encapsulate everyday experience in its respective cultures (Schwanenflugel, Blount, & Lin, 1991; Scarr, 1994). Thus, individuals from various cultures may well differ in their definitions of common words. Communities that are purportedly speaking the "same" language may use the same words to mean different things.[13]

The extensive work of Diaz-Guerrero and Szalay (1991) illustrates the different implications of common vocabulary use. They report on the different associations of African Americans, Whites, Mexicans, Puerto Ricans, Colombians, mainland Chinese, and Hong Kong Chinese for a wide array of com-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

monly used words, such as "friend," "love," "sex," "religion," "education," and "money," as well as an array of less commonly used words, such as "communism" and "capitalism." They present graphics and various similarity measures that reveal the degree to which ethnic groups differ in their associations for particular words. For example, the African American and White groups disagreed strongly in their responses to frequently used words such as "goals," "desires," "valuable," "justice," "progress," "society," and "class." On the other hand, African Americans and Whites agreed strongly on other terms, which happened to be words with low frequency of occurrence, such as "capitalism" and "communism." One should not conclude that these groups agree only when rarer words are presented as stimuli. In fact, African Americans and Whites were found to agree strongly on such commonly used words as "school," "father," and "mother." Nevertheless, Diaz-Guerrero and Szalay (1991) present provocative ethnic contrasts that provide a useful explanation of why any two cultural groups might differ in their responses to high-frequency words that often appear as *easy* analogy and antonym items on standardized tests, such as the SAT.[14]

It will be useful to encapsulate the above comments in a two-part hypothesis: (a) Performance on many *easy* verbal items is hypothesized to be highly dependent on the semantic sense of common words that are used in everyday conversation within a given community. Many *easy* verbal items tap into a more culturally specific content and therefore are hypothesized to be perceived differently, depending on one's particular cultural and socioeconomic background. Thus, the cultural and lexical ambiguity that African Americans are hypothesized to experience when responding to many *easy* verbal items offers one promising explanation for why they and other minorities do differentially worse on at least the *easy* analogy and *easy* antonym items; (b) *Hard* verbal items often involve rarely used words that are hypothesized to have fewer potential differences in interpretation across ethnic communities. Such precise meanings for rare words are probably most often encountered in classroom lectures and in textbooks; *hard* items are hypothesized to be oriented toward curriculum or achievement. Therefore, the pattern observed for *hard* analogy and antonym items among matched-ability African American and White peer groups is most likely a consequence of two forces: 1) *hard* items tend to embody less ambiguous vocabulary, and 2) the statistical fact that if a minority person gets a fixed score of 500 on the verbal SAT and has performed less well on the *easy* items than the White candidates, this mathematically implies that the minority individual must have performed better on the remaining *hard* items.[15]

The Rationale for Focusing on Hard-*Item Test Performance*

The purpose of the R-SAT score as presented in this article is to minimize cultural bias associated with *easy* verbal items. As such, our analysis eliminates the *easy* items and focuses instead on *hard*-item performance. Can one justify

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

focusing only on *hard*-item performance? I argue yes, because it already is a central assumption underlying computer-adaptive testing.[16] Specifically, computer-adaptive testing assumes that the best estimate of a candidate's proficiency can be determined from the highest level of item difficulty at which an examinee can reliably perform (Urry, 1977). Computer-adaptive testing focuses on difficult items. Our proposed R-SAT score invokes basically the same assumption: the ability of an examinee to perform well on the *hard* verbal items is a better indicator of true competence than is his/her performance on the *easy* items. Furthermore, I hypothesize that if there is a difference between an SAT score that combines performance on both *easy* and *hard* items and a revised R-SAT score that examines performance on just the *hard* half of the test, then the more accurate assessment of true proficiency lies with whichever score is greater. While this hypothesis is intended to benefit primarily the minority populations, we shall see evidence below that *all* students can potentially benefit to varying degrees.

Schaeffer et al. (1998) conducted a study comparing student performance on both a traditional paper-and-pencil (P&P) test and a computer-adaptive test (CAT), which selected items from a large pool of possible items. The items administered were different on the two tests. Students were randomly assigned to the study's two main conditions. Schaeffer et al. show that White examinees gain very little (from 505 to 507) on the verbal section when scores on the P&P and the CAT are compared. However, African Americans who take the CAT do significantly better than African Americans who take the P&P version. In fact, their verbal test score improves on average by about twenty points, from 371 to 390. Furthermore, African Americans' mean performance on their CAT quantitative scores improves by twenty-three points over their P&P scores.

A similar improvement on the verbal section of the CAT occurs, but to a lesser degree, for Hispanic Americans (7-point gain) and Asian Americans (12-point gain). For the quantitative sections, Hispanic Americans improve a substantial thirty points, while the Asian Americans improve by about forty points (Shaeffer et al., 1998, Table 5, p. 17).

Why do the verbal scores improve for minorities but very little for Whites? My explanation is straightforward. The CAT measures student proficiency by focusing on the hardest level at which the student can reliably perform; however, the paper-and-pencil version focuses on a scoring method that gives *easy* and *hard* items equal weight. One possibility to consider for reducing ethnic bias in the SAT could be to pursue use of the CAT for the SAT. However, there are technical and financial difficulties in pursuing the CAT. First, there are not enough computer terminals to handle the large volume of examinees that currently take the SAT, and second, there are financial problems in generating a very large item pool for which a testing organization can maintain security. For this reason, we focus our attention on revising the paper-and-pencil SAT.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Calculating the Revised-SAT

My recommendation regarding the SAT is to focus on *hard*-item performance, using the R-SAT score as an alternate estimate of a student's true verbal proficiency. Here I illustrate how to calculate the R-SAT scores for groups of African American individuals. Table 2 helps guide the explanation.

The first row in Table 2 presents all examinees (White and African American) who originally scored 200 on their SAT (the lowest possible score). All the White examinees are further examined for how well they responded to only the forty hardest items, and a percentage correct score is entered in the table; we see that 11.9 percent of the forty hardest items were correct. In other words, of the forty *hard* items, 11.9 percent is the *average* percentage correct across all Whites who scored 200. A similar calculation was done for all African Americans whose original score was 200; we see that 13 percent of the forty hardest items were correct. One notes that African Americans got more *hard* items correct than their "matched" cohorts among the White examinees. Since I assert that *hard*-item performance is a better indicator of true ability, it is reasonable to ask a further question: since the African Americans got more *hard* items correct, how far down in the ability scale would one have to go in order to find a similar *hard*-item performance level among the White examinees? One can see that Whites who got an original SAT score of 210 got 12.9 percent of the *hard* items correct; this 210 is the White entry that most closely approximates the African American *hard*-item performance at 200. So, one assigns a "gain" score of 10 points to the African American's original score of 200, and enters the R-SAT score of 210 (200+10 = 210) in the right-hand column. The next entry for 210 shows again that African American examinees outperform matched-ability Whites on the *hard* items, 14.6 percent versus 12.9 percent.

How far down the White column must one go in order to find a *hard*-item performance level that most closely approximates the African American performance of 14.6 percent? One must go down four steps, to 250, where the White performance exactly equals the African American *hard*-item performance. Hence, African American examinees who originally earned 210 on the SAT can now be assigned a 40-point gain, which is added to their original SAT score of 210 to yield an R-SAT score of 250, which is entered in the right-hand column. In general, one can see that African Americans do better on *hard*-item performance than Whites with the same total SAT scores.

Another pattern that emerges from Table 2 is that the maximum increases for African Americans occur early in the table. As shown, the maximum gain score for African Americans is 60 points, which occurs at the SAT levels of 230 and 240. As one moves down the table — as the original SAT score gets larger — the gain scores begin to trail off to about 10 points. And as one approaches even higher scores (not shown in the table), the results typically oscillate around zero. The reason that the gain scores are larger at low SAT lev-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 2 *Percentage of Correct Responses to the 40 Hardest Verbal Items Used in an Early SAT Test**

*Original SAT Score for 85 Verbal Items***	*White Examinees % Correct (40 hardest)*	*Black Examinees % Correct (40 hardest)*	*Gain Score for Black*	*Average Reestimated R-SAT Score for Black*	*Bl-Wh % Diff.****
200	11.9%	13.0%	10 pts	200+10 = 210	1.1%
210	12.9%	14.6%	40 pts	210+40 = 250	1.7%
220	13.2%	15.6%	60 pts	220+40 = 260	2.4%
230	14.2%	16.1%	60 pts	230+60 = 290	1.9%
240	14.5%	16.5%	50 pts	240+50 = 290	2.0%
250	14.6%	17.2%	50 pts	250+50 = 300	2.6%
260	15.1%	17.6%	40 pts	260+40 = 300	2.5%
270	15.2%	18.2%	40 pts	270+40 = 310	3.0%
280	15.9%	18.4%	30 pts	280+30 = 310	2.5%
290	16.1%	19.0%	40 pts	290+40 = 330	2.9%
300	17.4%	19.2%	30 pts	300+30 = 330	1.8%
310	18.3%	20.5%	40 pts	310+40 = 350	2.2%
320	18.1%	21.1%	40 pts	320+40 = 360	3.0%
. . .					
470	33.9%	35.3%	10 pts	470+10 = 480	1.4%
480	35.6%	36.4%	10 pts	480+10 = 490	0.8%
490	36.5%	37.7%	10 pts	490+10 = 500	1.2%
. . .					
640	68.0%	69.2%	10 pts	640+10 = 650	1.2%
. . .					
800	100%	100%	0 pts	800+00 = 800	0.0%

* This test is designated by the symbol OB023 (or 3CSA2) in the ETS file system.

** The values entered for this form use the R+W formula, which is the formula recommended by Kulick and Hu (1989); however, see Appendix B for further comments.

*** This column subtracts the White percentage correct from the African American percentage correct.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

els is that the percentage difference between the two ethnic groups in *hard*-item performance is greater at these low SAT levels. It seems likely that the greater gains that occur at these lowest SAT levels are due to the possible increased representation of lower socioeconomic status (SES) at these levels, but a separate analysis of this possibility has not been conducted. Table 2 shows the percentage difference between the White and African American examinees for each SAT score level — it begins with a 1.1 percent difference for the 200 SAT level, increases to about 3.0 percent difference around 270, and then begins to trail off to about 1.2 percent or less for higher SAT levels. For the particular calculations entered in Table 2, the average gain score for African Americans equals 30.08 points. This average group gain by itself does not seem particularly substantial; however, when individual R-SAT scores are calculated, the gains can be considerable.

Estimating Individual R-SAT Scores: An Initial Illustration

Returning to Table 2, one can see that the maximum gain shown is sixty points for the group of African Americans who originally scored 220 and 230. However, if the group has gained sixty extra points, this implies that there should be many individuals who have gained even more (i.e., all those individuals above the group mean). Since we are interested primarily in those individuals who score even higher on the *hard* items (because of the college admission implications), it will be useful to attach a separate R-SAT for each individual examinee.

For purposes of illustration, the logic of the procedure for individual gain scores will be identical to the way we assigned gain scores for groups of individuals; however, later in this report, a more robust technique will provide the final recommended procedure for facilitating R-SAT assignment. For example, if we know an African American examinee originally got a score of 220 but got 18.1 percent of the *hard* items correct, then, as Table 2 shows, this person can be assigned an R-SAT score of 320, a gain of 100 points. Similarly, for any other White or African American examinee, once the White group's percentage of correct responses to the *hard* items has been calculated for each original SAT level, as in Table 2, the data entered provide all the information needed to reassign individual R-SAT scores.[17] The following section will present a more in-depth estimation of individual R-SAT scores.

The Revised-SAT for High-Scoring Individuals

I will begin by focusing on individuals at each SAT score level who have yielded the highest performance on the forty hardest verbal items.[18] Table 3 presents the gains in R-SAT scores for these students. These top-scoring individuals are selected to illustrate that individual R-SAT scores can show very large gains for individuals whose original SAT scores are often quite low.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Many of these exceptional individuals, I maintain, might have been considered for admission into colleges that had high SAT cut-off scores among other criteria for acceptance.

The first part of Table 3 presents the highest scoring individuals at each original SAT level from 200 through 390. Notice that there is an African American student who scores at R-SAT = 600, even though his/her original SAT score was only 290. This student's gain score is 310 points — an astonishingly large reassessment of his/her scholastic skills.[19]

In general, the top-scoring African American student at each original SAT score level gains using the R-SAT method of scoring his/her verbal skills. This is borne out by examining the average scores (see row labeled "Average 200–290" in Table 3). The average top-scoring African American student's R-SAT score is 543, while the top-scoring White student in the same row averages an R-SAT score of 514, which is still impressive but not as high as the African American student's score.[20]

What does the R-SAT score mean? What does it mean for individuals and the test itself? Since the R-SAT is defined here for only the *hard* half of the verbal test — which is basically consistent with CAT assumptions — would it have made a difference if we also had assigned a second R-SAT score using just the *easy* half of the verbal test? Consistent with my two-part hypothesis concerning cultural bias presented earlier, I generally interpret the difference between an individual's R-SAT score and his/her regular SAT score (i.e., R-SAT minus SAT) for the *hard* half as a measure of the degree to which this individual's cultural background diverges from White, middle-class culture.

Because a general symmetry (except at the extreme low end of the SAT scale) can be expected between how well any disadvantaged student of low socioeconomic status or minority status overperforms on the *hard* half (vis-à-vis his/her regular SAT score) and how poorly the same student underperforms on the *easy* half, it should be clear that a separate measure of performance on only the *easy* items in most cases would amount to a measure of the same underlying concept: departure from White, middle-class culture.[21]

Social Class, Language Background, and Test Bias

White students comprised 78 percent of the test-taking population in the 1980s.[22] Of this group, 54 percent come from families whose income is $50,000 or greater; only 12 percent come from families earning $20,000 or less. The question is whether family income differences among Whites (and associated differences in education level) lead to a sufficiently different set of cultural values and expectations so that one might expect to find a significant SAT test-bias effect emerging from analyses that contrast subgroups of Whites with different income and/or education levels. If so, then one can ex-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 3 *The Best Scoring Individual at Each SAT Score Level with His/Her R-SAT Score*

Original SAT Score	*Form OB023 White Gain*	*Form OB023 White R-SAT*	*Form OB023 Black Gain*	*Form OB023 Black R-SAT*	*Bl-Wh***
200	220	420	250	450	30
210	230	440	290	500	60
220	240	460	300	520	60
230	270	500	270	500	0
240	300	540	300	540	0
250	280	530	310	560	30
260	300	560	330	590	30
270	290	560	310	580	20
280	290	570	310	590	20
290	270	560	310	600*	40
Average 200–290	269	514	298	543	29
300	310	610*	310	610*	0
310	300	610*	290	600*	–10
320	290	610*	340	660*	50
330	290	620*	330	660*	40
340	290	630*	310	650*	20
350	300	650*	340	690*	40
360	280	640*	320	680*	40
370	280	650*	290	660*	10
380	260	640*	290	670*	30
390	270	660*	300	690*	30
Average 300–390	287	632*	312	657*	25
Average 400–490	233	681*	255	700*	19
Average 500–590	189	734*	219	764*	28

* The asterisk indicates a student whose original SAT score is below 600 but whose R-SAT score is 600 or higher. All SAT scores (as well as R-SAT scores) are uncentered.

** The column labeled "Bl-Wh" indicates that the White Gain score is subtracted from the African American Gain Score. The data represent the most conservative R+W+O+NR scoring formula (see Appendix A).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

pect not only Blacks but also some White students to benefit from application of an R-SAT score.

A brief survey of the research literature provides evidence that, among Whites, there is often a cultural difference associated with education and/or income level. For example, Hall and Freedle (1975) reported several significant differences in language use associated with social-class differences for Whites and Black students. In addition, Cook-Gumperz (1973) summarizes significant White differences in language use and childrearing practices in Great Britain as a function of social class.

I have taken a recent form of the SAT (QSA01) to provide empirical evidence favoring the R-SAT score. From Form QSA01, I selected 263 Whites from the highest income ($70,000 or more) and parental education levels (college degree or higher). Their responses to the analogy items were contrasted with 370 lower-income Whites (earning $20,000 or less) whose parents had less education (a high school diploma or less). The six easiest analogies yielded a DIF score of –.031, which indicates that the lower-income Whites performed less well than the higher-income Whites. For the hardest analogies the DIF score was +.016, indicating that the lower-income Whites performed slightly better than the high-income Whites. This yields the familiar bias pattern — as observed with African Americans — indicating, in this case, socioeconomic bias against the White students from low-SES backgrounds.[23]

While it is somewhat flawed in its design, a second example provides a similar result. I took another recent SAT form (QSA09) and defined two additional White subgroups based on slightly higher SAT scores. Both groups scored between 300 and 390 on the traditional SAT. One group of 516 Whites was from poorer income levels ($30,000 or less) and English was not the sole preferred language. This group was compared to White students from all income levels (but with a predominance of higher incomes, inasmuch as 88% exceed the incomes of the disadvantaged group) and for whom English was the preferred language (N = 56,672). The easiest five analogies yielded a DIF value of –.028, which indicates that the disadvantaged Whites performed less well on these easiest analogies. The five hardest analogies yielded a +.014, indicating that the disadvantaged Whites performed slightly better on the hardest analogies. Again, we observe a familiar bias pattern. The existence of this bias pattern further justifies the calculation of an R-SAT score for Whites, as well as for any of the ethnic groups.

The Implications of an R-SAT Score for High-Scoring Individuals

Below I hypothesize that receiving a revised verbal score of at least 600 (actually, 670 on the recentered scale) would be sufficiently meritorious to interest many colleges in an applicant who received such a score. My choice of a score of 600 is not totally arbitrary. One can see by examining data on the College Board website that students whose high school grade point average

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

is between 97 percent and 100 percent receive an average SAT verbal score of 610 (recentered). Colleges undoubtedly differ in the weight they attach to SAT scores.[24] A score of 600 also reflects a level of test performance that only about 5 percent of the test-taking population receives, using the normal SAT scoring procedures.

Table 4 is crucial for understanding the implications of this article for college admissions. Even at the lowest scores of 200–290, there are a few students who, by virtue of the R-SAT estimation procedure, will likely increase their chance of admission to a select college. This is primarily true for several of the low-scoring African American students. For example, for Form 4I, in the original SAT 200–290 range, nine African Americans out of 3,605 earned 600 or more points on their revised R-SAT score.[25]

One can see the frequency of probable increases in select college admission for African Americans and Whites as one moves up the original SAT scale, so that, by the time we consider students originally getting 500–590 on their SAT, about 21 to 23 percent of the African Americans and Whites are now assigned very high R-SAT scores. For Form 4I, without the R-SAT scores added (this is the row designated 600–800), one sees that only 0.72 percent of the original African American sample will likely go to a select college; however, after adding in the R-SAT scores, this percentage of probable select college admissions increases to 2.46 percent. On Form 4I, for the original SAT scores (600–800), only 5.19 percent of Whites are likely to go to a select college, but after adding in their R-SAT scores this percentage increases to 10.20 percent. These results are basically the same for the other form (OB023).

By calculating the percentage increase in the scores of high-scoring African American and White students, we can see that African Americans can benefit proportionately more than Whites from the new way of scoring. For Form 4I, we said that 0.72 percent of African American students will potentially gain admission to select colleges using the original SAT scores (600–800); by adding in the R-SAT, this percentage increases to 2.46 percent. The ratio of 2.46/0.72 x 100 represents *an increase of 342 percent.* For Whites, the increase from 5.19 percent to 10.20 percent represents *an increase of 196 percent.*

A similar dramatic increase in potential African American enrollments occurs for the other test form (OB023) presented in Table 4. For this second form, the increase goes from 0.64 percent of African Americans being admitted (because they get scores of 600–800) to a grand total of 2.14 percent (obtained after adding in the African American R-SAT scores of 600 or greater), which represents an increase of potential African American enrollments of 334 percent (2.14/0.64 x 100 = 334). For Whites, the comparable increase is from 6.56 percent to 12.41 percent, an increase of 189 percent. These figures agree closely with what was found for Form 4I. The implication of these results is that more African Americans and disadvantaged Whites will likely qualify for admittance into select colleges.[26]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 4 *The R-SAT Score*

Original SAT Score	*Form 4I Freq.* Black with R-SAT of 600+*	*Form 4I Freq.* Whites with R-SAT of 600+*	*Form OB023 Freq. Black with R-SAT of 600+*	*Form OB023 Freq. Whites with R-SAT of 600+*
200–290	9 (3605; 0.25%)	0 (2401; 0.00%)	1 (3595; 0.00%)	0 (2920; 0.0%)
300–390	12 (3687; 0.33%)	13 (9996; 0.13%)	10 (2761; 0.36%)	10 (12564; 0.08%)
400–490	57 (1779; 3.20%)	328 (11645; 2.82%)	35 (1071; 3.27%)	358 (14215; 2.52%)
500–590	88 (419; 21.00%)	1236 (5809; 21.28%)	71 (336; 21.13%)	2037 (8767; 23.23%)
Subtotal	166 (9490; 1.75%)	1577 (29851; 5.28%)	117 (7763; 1.51%)	2405 (38466; 6.25%)
	Form 4I Freq. Black with Regular SAT of 600+	*Form 4I Freq. Whites with Regular SAT of 600+*	*Form OB023 Freq. Black with Regular SAT of 600+*	*Form OB023 Freq. Whites with Regular SAT of 600+*
600–800	69 (69; 100%)	1635 (1635; 100%)	50 (50; 100%)	2702 (2702; 100%)
Subtotal	69 (9559; 0.72%)	1635 (31486; 5.19%)	50 (7813; 0.64%)	2702 (41168; 6.56%)
	Form 4I R-SAT + SAT (600 & higher)	Form 4I R-SAT + SAT (600 & higher)	Form OB023 R-SAT + SAT (600 & higher)	Form OB023 R-SAT + SAT (600 & higher)
Grand Total	235 (9559; 2.46%)	3212 (31486; 10.20%)	167 (7813; 2.14%)	5107(41168; 12.41%)

*These frequencies also include those few maximum scoring African Americans and Whites (i.e., the best-performing individuals at each score level with respect to the forty most difficult verbal items). Within parentheses, I list the number of examinees within a designated SAT range followed by the percentage of examinees whose R-SAT score equals or exceeds 600. All these calculations for African Americans and Whites used the conservative R+W+O+NR formula (see Appendix A).

Bias in the Current SAT Format

The newest version of the SAT that has been used over the last several years is different from the older versions of the test that I employ in most of my analyses. The current SAT may have been modified to include different numbers of each verbal item type, but there is continued empirical evidence that ethnic bias is present. I offer this evidence in Table 5.

In Table 5, we see that there are significant correlations between 116 new analogy items and the DIF for the current SAT format. The correlations for African American and Hispanic Americans, .496 and .555, respectively, are

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 5 *Correlations between DIF and Item Difficulty for 116 New Analogies Items and 90 New Sentence Completion Items for Three Ethnic Groups*

Ethnic Group	*Analogy Correlation (n = 116)*	*Sentence Completions Correlation (n = 90)*
African American	.496 *** +	.275 **
Hispanic American	.555 ***	.263 *
Asian American	.296 **	.495 ***

* Significant, p < .05 (2-tailed)

** Significant, p < .01 (2-tailed)

*** Significant, p < .001 (2-tailed)

+ A positive correlation between DIF and item difficulty implies that the *easy* items are differentially responded to more poorly by the given ethnic group while, at the same time, the *hard* items are differentially responded to better than matched White examinees.

quite large. The correlation for Asian Americans, .296, is smaller, but significant.

In general, a positive correlation means that each minority group is responding better to the *hard* verbal items and worse to the *easy* analogy items than matched-ability White examinees. This is exactly the pattern reported by Freedle and Kostin (1990). Similarly, on ninety new sentence-completion items the correlations for the three minority groups are significant, but generally not as large as those produced for the analogies (the Asian Americans are an exception).[27]

Quantitative SAT Items

Kulick and Hu (1989) found that ethnic bias was present for a sample of 540 quantitative SAT items administered from 1986 to 1987. The familiar bias pattern is found: *hard* items are differentially responded to better by the minority groups in comparison with their matched-ability White peers. The *easy* quantitative items show the converse effect, reporting a correlation of .35 for African Americans between math item difficulty and DIF magnitude, .33 for Hispanic Americans, and .22 for Asian Americans. Their results did not distinguish between students who report English as their best language and those who report equal competence in English and another language. In addition, Kulick and Hu did not include students whose preferred language is not English. Nevertheless, I find that their reported bias regarding math items can be replicated with more recent SAT data, as illustrated in Table 6.

In Table 6, we present new results for forty-eight quantitative items from one recent SAT test form (QSA09). One can see a similar pattern of bias; for

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 6 *Differential Ethnic Effects for 48 SAT Quantitative Items from a Recent Test Form (QSA09) for Examinees with SAT-M Scores between 200–290*

Ethnic Group	*Average DIF for 24* Easy *Quantitative Items*	*Average DIF for 24*** Hard *Quantitative Items*
African American (English best language) (N = 3,903)	–2.04	+1.08
African American (English not best)* (N = 362)	–2.38	+1.38
Hispanic American (English best language) (N = 746)	–1.46	+0.54
Hispanic American (English not best)* (N = 1,047)	–1.83	+1.71
Asian American (English best language) (N = 261)	–1.38	+1.08
Asian American (English not best)* (N = 335)	–0.79	+1.17
AngloAmerican (English not best)* (N = 278)	–1.58	+1.33

* Two categories of respondents were included here, those who checked "other language" as their best language and those who checked "English and another language about equal."

** Two middle-difficulty items were dropped out of a total of fifty quantitative items in defining the *easy* half versus the *hard* half. The White comparison sample for all six ethnic group comparisons consisted of 5,147 White students whose primary language was solely English and who themselves scored in the 200–290 range on their SAT-M score.

example, the African American group (N = 3,903) performs on average about two percentage points below (–2.04) matched-ability Whites (N = 5,147) on the twenty-four easiest math items, and on average about one percentage point above (+1.08) the same matched-ability Whites on the twenty-four hardest math items. For African Americans whose sole preferred language is not English (N = 362), the DIF bias pattern increases, yielding a DIF score of –2.38, or more than two percentage points below matched-ability Whites on the *easy* math items, and +1.38, or more than one percentage point above matched-ability Whites on the *hard* items.[28]

One can observe a similar bias pattern for Hispanic Americans. The DIF bias increases when the sample is comprised only of those students for whom English is not their sole preferred language. Asian Americans show the same

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

bias patterns with one exception; they fail to show an increase in their negative DIF on the *easy* items when English is not their sole preferred language.

Of further interest is the White sample (N = 278) for whom English is not the sole preferred language; these individuals were contrasted with the White sample used in all the previous DIF comparisons (N = 5,147): those Whites for whom English is the sole preferred language. This special sample of Whites produces a fairly substantial average negative DIF of –1.58 on the twenty-four *easy* math items, and a comparatively strong average positive DIF of +1.33 on the *hard* items.

This result for contrasting two White samples shows that it is not ethnicity per se that is producing these DIF patterns; the source is, rather, any index that identifies a group as sharing a persistent environment that differs from the White majority English speakers. A linguistic difference in that sense can be thought of as a type of cultural difference; membership in any such group creates some degree of cultural mismatch that reveals itself in the typical DIF pattern of *easy* items and *hard* items that produce quite divergent results.

Carlton and Harris (1992) described the types of quantitative items that yield better African American performance (positive DIF) vis-à-vis their White peers. They showed that quantitative items that were rated as more "textbook like" or more "abstract" yielded positive DIF scores for African American examinees; this generalization also applied to the Hispanic American and Asian American groups.[29] To help explain why the quantitative items produce some of the DIF bias patterns, we can apply the cultural familiarity hypothesis, as described earlier for the verbal DIF findings. The *hard* math items are more likely to contain both rarer and more abstract concepts learned strictly in the classroom, and therefore should be less sensitive to cultural background. In contrast, the *easy* items are likely to contain more common vocabulary terms whose exact meaning is sensitive to cultural background. If further empirical work confirms this hypothesis, it should be possible to extend the same two-part hypotheses to the quantitative items that I used to explain our earlier ethnic differences for the verbal items.

Whatever final cognitive reasons explain the differential ethnic responses, the practical solution for mitigating math bias from the SAT should be clear. The Educational Testing Service (ETS) and the College Board should compute a revised quantitative SAT score to supplement the regular quantitative SAT score. Further, ETS and the College Board should study the reliability and predictive validity of the SAT by doing a detailed study of the correlation with freshman grades.

Bias in Two Advanced Placement Tests: Implications for the SAT II

This section focuses on DIF results for the multiple-choice section of two Advanced Placement (AP) Tests: U.S. history and biology.[30,31] I argue that the bias shown for the multiple-choice sections of these two AP tests might occur

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

for the SAT II Achievement Tests, inasmuch as the structure and content of the multiple-choice items are similar.[32]

In Table 7, I present the DIF results for these two AP tests. In 1994, the multiple-choice part of the AP biology test consisted of 120 questions that ranged from very *easy* (90% of students respond correctly) to very *hard* (10% of students respond correctly). Table 7 reveals that there are highly significant DIF correlations for each of the three minority groups. The largest ethnic DIF correlation (.331) is for Asian Americans; this correlation indicates that these students (as well as the other two minority groups) generally do worse on the *easy* items and better on the *hard* than do matched-ability White students.

The results for the U.S. History Advanced Placement Test (administered in 1999) also yield highly significant ethnic differences, with the largest DIF entry (–.342) for African Americans.[33] There are a total of eighty history questions for this AP test. As with other tests, African Americans respond differentially better to *hard* items than to *easy* items when compared to matched-ability White students. Hispanic Americans and Asian Americans show the same bias pattern as African Americans.

What is most surprising about these data is that these test questions represent material selected from a preestablished reading list. All students knew beforehand what material was to be tested. Yet, minority students still responded differentially worse to *easy* items and differentially better to *hard* items. Such results suggest that there is something about the structure and content (e.g., the semantic and linguistic features) of the items that are contributing to these systematic effects. *Easy* items probably use cues that introduce sufficient ambiguity of interpretation to cause all minority groups to perform differentially more poorly; exactly the converse occurs on the *hard* items. Whether the hypothesized ambiguity is traceable to common vocabulary terms has not yet been analyzed; it is therefore unclear to what degree the hypothesis presented to explain the verbal SAT results necessarily extends to explaining the findings for these tests. Suffice it to say, minority students lose enough points on the *easy* items so that it is very difficult, in fact impossible, for them to regain sufficient ground when responding to the *hard* items to show their true ability. The way around this problem, as before, is to calculate a revised score following the logic employed in the earlier sections of this article.

The SAT II Achievement Tests (the subject-area tests) are similar in content and form to the multiple-choice sections of the AP tests; therefore, I fully expect the same ethnic bias patterns to emerge once the SAT II is subjected to DIF analyses. These patterns are important, given that the University of California system is currently considering substituting the SAT II tests for the SAT I. Because scores are crucial in the college selection process, ensuring that the SAT II tests are unbiased is critical.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 7 *Ethnic Bias on Two Advanced Placement Tests, U.S. History and Biology: Correlation between DIF and Item Difficulty*

Ethnic Group	*Advanced Placement Test*++ *U.S. History (1999)*	*Advanced Placement Test*++ *Biology (1994)*
African American	−.342** (.401**), N = 80	.272** (.339**), N = 120
Hispanic American	−.315** (.399**), N = 80	.282** (.358**), N = 120
Asian American	−.266* (.292**), N = 80	.331** (.509**), N = 120
Male/Female	−.003ns (.027ns), N = 80	.004ns (.001ns), N = 120

* Significant, p < .05 (2-tailed)

** Significant, p < .01 (2-tailed)

ns = not significant, p > .05 (2-tailed)

++ The U.S. History test involved equated delta as the measure of item difficulty; the Biology test used Percentage Correct as the measure of item difficulty — hence the reversal of algebraic sign. The correlation enclosed within parentheses represents a slightly more sensitive test of ethnic bias called the Mantel-Haenszel, which is now used at Educational Testing Service as a replacement for the Standardized DIF statistic (see Holland & Thayer, 1988).

The Final Operational Method for Estimating Revised-SAT Scores

Up to this point I have used a method of estimation that relies on comparing African American and White examinees (and other groups) who have been matched on their SAT performance. This, of course, is the essence of the DIF technique. There are many insights to be gained from this approach, but it also has several limitations. Most important, it assumes that students will always identify their respective ethnic group membership on their SAT background information sheet. In every test administration, a small number of examinees invariably fail to identify themselves. Therefore, it is impossible to calculate an R-SAT score (for math and/or verbal) for all individuals.

Fortunately, there is another new and more direct way to estimate the R-SAT score — one that ignores aspects of race, gender, and class. It is as follows: 1) one simply scores only the *hard* half of the SAT test for every individual; 2) from these individual scores one forms a single distribution of scores; 3) the raw scores (correct, incorrect, omissions, number not completed) are then optionally corrected for guessing;[34] and 4) a new scaled score from 200 to 800 is defined. All individuals who do well on only the *hard* items, regardless of who they are and what their background is, will get a high score, and those who do poorly will get a low score.

The above approach is ideal. Since I no longer have access to the ETS data tapes, I will use an indirect approach to estimate what the total distribution of these R-SAT scores might look like. I will again have to resort to the data at

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

hand (the SAT data analyzed by Freedle & Kostin, 1988, 1990). This data is admittedly flawed in the sense that not all African Americans and not all Whites have responded to the background questionnaire, but it should provide a close approximation. For that reason, it will suit my purpose of showing the likely impact of the R-SAT score on the distribution of African American and White performance, respectively. If I combine all White and African American students into a single distribution, this typically represents the great majority of people (87%) who take the SAT.[35] Therefore, the combined distribution of White and African American students' performance on just the *hard* half of the test will, for the purposes of this section, be my best approximation to the total population of test-takers. To see how well African Americans and Whites will do within this new distribution of scores for the *hard* items, I focus on just one test form (4I) for this demonstration, as displayed in Table 8.

Table 8 presents the overall mean and standard deviation for (a) self-identified Whites, (b) self-identified African Americans, and (c) the total African American and White population, as based on the conservative R+W+O+NR formula for R-SAT score estimates (see Appendix A). To contrast African American and White performance with this test, I use a statistic called "sigma" that expresses the difference between the two ethnic populations in terms of a standardized difference calculation.

A small sigma indicates that the test performance of two ethnic groups is very similar; a large sigma indicates that the two groups are dissimilar. Previous research has reported that Whites and African Americans differ by about one standard deviation on the SAT as currently scored; that is, sigma typically equals about 1.00 (e.g., Jensen, 1980; Herrnstein & Murray, 1994), a value that has been considered to be quite large. I believe the sigma can be significantly reduced.[36]

Table 8 shows that the traditional SAT results for all the verbal items yield a sigma of .90, which is very close to the typical value of 1.00 reported in the literature. The second row shows that when only *hard* verbal items are examined, the two ethnic groups move much closer together so that the sigma now equals .62; this represents a reduction of nearly one-third of a standard deviation.[37] It is clear that the perspective taken by this report seriously calls into question the empirical basis for the racist conclusions drawn by such well-known authors as Jensen (1969, 1980) and Herrnstein and Murray (1994). Issues of further relevance to this matter are taken up below.

Although I no longer have access to the raw data, it is evident that the sigma difference between Whites and African Americans for all eighty-five test items is approximately 1.00 (i.e., one standard deviation difference). If the sigma for only the forty hardest items is reduced to about .60, this implies that had a separate sigma been calculated for only the easiest items, the value would almost certainly be greatly in excess of 1.00. Such a result would clearly have shown that African Americans and Whites differ more dramati-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 8 *A Comparison of White and African American Mean Performance (Using the Traditional SAT and the R-SAT for Test Form 4I)*

Type of Items Analyzed	*Mean White Performance*	*Mean Black Performance*	*Standard Deviation of Whites plus Blacks*	*Sigma**
Easy and *Hard* Verbal Items**				
(85 items)	428.89	336.71	102.14	.90
(R+W+O+NR)				
Hard Verbal Items***				
(40 items)	30.95%	21.98%	14.45	.62
(R+W+O+NR)				

*Sigma is calculated as follows: The African American mean is subtracted from the White mean and the result is divided by the standard deviation of the combined African American and White sample—for example, for 85 items, [(428 –336.71)/102.14] = [92.18/102.14] = 90. This calculation indicates that, for the typical SAT method of scoring, the African American population is almost one standard deviation below the White mean.

**The sigma for the hard and easy items (N = 85) reflects the correction for guessing that is normally applied to SAT scores. Score values range from 200 to 800.

***The sigma for expressing the difference between the mean performance of the African American and White examinees for only the *hard* items (N = 40) uses the percentage pass score. It is not expressed on the 200 to 800 scale. Also, these values do not apply the correction for guessing. A separate analysis of similar data for only the *hard* items indicated that the correction for guessing does not strongly affect the size of the sigma. In this data set the sigma varied from .68 without the correction and was .66 after the correction. I conclude that sigma for only the *hard* items indicates that the African American and White examinees are much closer in comparison with the standard calculation that includes both *easy* and *hard* verbal items.

cally on the *easy* items while at the same time are more similar on the *hard* items. Such a finding would be consistent with Flaugher and Schraeder's (1978) findings. These two findings imply a large cultural difference for the *easy* items and a much smaller cultural difference for the *hard* items.

Thus, if we regard *hard*-item performance as a more accurate indicator of true underlying ability, then the measured difference between White and African American SAT examinees is shown to be substantially reduced. Had we been able to control for the additional negative effects of family income, quality of education, and prior practice (see Powers & Rock, 1999), it is predicted that the calculated sigma would be even further reduced, perhaps approaching zero — a result that would not surprise many biologists (Gould, 1995), anthropologists (Gumperz, 1982), and linguists (Labov, 1975) who have critically commented on the testing literature regarding ethnic differences. In short, the methodology introduced in this article may provide yet another approach, given sufficient future interest, toward affirming the essential intellectual equality of all races.[38]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Critiquing Old and New SAT Designs in Terms of Ethnic Bias

In this section I critique the old SAT format, provide a critique of the current SAT format, suggest additional strategies for analyzing reliability and validity in both formats, and briefly survey a few other formats, including the one most likely to be implemented by the College Board.[39] Each format shows a different degree of racial bias. For each design, I consider how an R-SAT or a variant of the R-SAT might ameliorate any bias that is found.

Reliability and Validity of Computing the R-SAT Score Using the Old SAT Format

We already have seen ample evidence that the old SAT is ethnically biased in both its verbal and quantitative sections (Freedle & Kostin, 1990, 1997; Kulick & Hu, 1989). The most troublesome verbal items in the old format are the *easy* analogies and the *easy* antonyms. To a lesser degree, the *easy* sentence completions are also troublesome. I have shown above how to generate at least one R-SAT score by focusing on the forty hardest items. I have argued above that the R-SAT is a measure (actually, the individual's R-SAT score minus the individual's SAT score) of the degree to which each individual deviates from his or her matched-ability White peers, and thus by extension from White, middle-class culture.

A cruder way of describing the R-SAT score is merely to say that it is what remains after eliminating the forty-five easiest items. An early study by Flaugher and Schraeder (1978; also see Angoff & Ford, 1973) actually showed that if only the twelve easiest verbal and fifteen easiest math items were discarded, the difference in mean SAT scores would diminish between African American and White students. They also noticed that removing these *easy* items had little effect on test reliability. This early work is therefore encouraging with respect to our R-SAT score — the real question is whether test reliability for the verbal would be greatly diminished if all forty-five of the easiest verbal items were discarded.[40]

How strongly does a student's R-SAT score correlate with his/her freshman college grade average? It is not possible to conduct a rigorously accurate estimate of validity, since many people who did not attend select colleges might have matriculated at such schools if their R-SAT scores had been used in the admissions process. Nevertheless, an attempt should be made to study these old data and compute a validity correlation. Is the validity coefficient (the correlation) higher between R-SAT and grades than the validity coefficient between regular SAT and grades?

Before I leave this aspect of the validity of the R-SAT score for the old SAT, I need to deal with the issue concerning different degrees of predictive validity for minority versus White populations. The reader will remember that Gould (1995) called attention to the fact that there are two general types of test bias: cultural and statistical. By statistical bias he indicated that a test

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

would be biased if two students who get the same final score on a standardized test nevertheless perform differently in some other setting, such as freshman college grades. Actually, Ramist, Lewis, and McCamley-Jenkins (1994) show that the SAT overpredicts the freshman college grades of African Americans when compared with those of White students. If we accept this premise, it would be reasonable to ask whether the R-SAT would still overpredict their freshman grades, just as the regular SAT does? Not necessarily. While many African Americans who get very high SAT scores will necessarily get high R-SAT scores, we have seen in Tables 3 and 4 that many low-scoring SAT individuals (African American and White) turn out to have quite high R-SAT scores. In other words, the distribution of individuals who might qualify for college admission under the traditional SAT will be substantially different from the population of individuals who qualify using their R-SAT scores. (Remember, the number of African Americans who get a high score of 600 or more increases by almost 400 percent over the number of African Americans who get 600 or higher using the traditional SAT score.) Therefore, it is an open question whether the predictive validity studies as reported by Jensen (1980) for the traditional SAT score necessarily apply to the R-SAT scores. The empirical work has yet to be done.

Reliability and Validity of the Current SAT Format

A logical new format to help reduce ethnic bias — as suggested by the empirical findings of Flaugher and Schraeder (1978), Angoff and Ford (1973), and Freedle and Kostin (1987, 1990) — would have been to increase the number of *hard* antonyms, *hard* analogies, and *hard* sentence completion items (and perhaps to reduce the number of reading items), since the earlier work all showed that the mean performance of African American and White examinees is more similar on the *hard* items (excluding reading items). Instead, based in part on the recommendations of an ETS research committee, the new format basically dropped all antonyms and added more reading items. Is this current format conducive to less ethnic bias and hence more minority enrollments? The significant DIF results reported above showing persistent ethnic bias for quantitative and verbal items (excluding reading) in the current SAT suggests that the answer is "probably not."

An R-SAT score calculated for the new format will be less reliable simply because there are now even fewer total *hard* items wherein the minority populations can demonstrate their true verbal competence (i.e., only several *hard* analogies and *hard* sentence completions are present in the current SAT). An attempt to increase reliability of the R-SAT by simply adding the *hard* reading-item performance to the computation would be misleading. Separate evidence (using the older SAT forms) shows a sharp increase in the gain score for African Americans when the mean R-SAT is recomputed after eliminating all the *hard* reading items, thereby leaving just the *hard* analogies and *hard* antonyms. Thus the R-SAT verbal construct seems strongly depend-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ent on measuring especially the *hard* analogy and antonym performance; boosting R-SAT reliability by simply adding in the *hard* reading items is therefore a questionable tactic.

As with the old SAT, predictive validity for the R-SAT (the correlation between freshman grades and R-SAT) is compromised for the current SAT simply because the R-SAT for the current test has never been calculated or reported to college admission officers for purposes of student admission. Therefore, if an R-SAT score is ever calculated for individuals (an unlikely prospect, given that the SAT is to be revised yet again), a retrospective validity study for the current SAT is likely to be tentative simply because those students who might have qualified for admission based on their unreported R-SAT score will at best have been admitted to less selective schools. Nevertheless, even with this admitted limitation, an interesting hypothesis can be advanced.

If R-SAT is calculated for the current SAT, there are two groups of disadvantaged students that should be distinguished; let us call them Group I and Group II. Both Groups are assumed to be admitted into less selective colleges. Group I consists of those disadvantaged students whose R-SAT verbal scores are much higher than their regular SAT verbal scores (for simplicity, I am considering just the verbal R-SAT here). In contrast, Group II consists of those disadvantaged students whose R-SAT scores are less divergent from their regular SAT scores. If we contrast the grades earned by students in Group I and Group II, it seems reasonable to predict that those in Group I whose R-SAT scores are much higher than their regular SAT scores should do significantly better in many of their college courses than students in Group II. If this prediction is borne out through empirical research, R-SAT scores do have predictive validity and should be reported (along with the regular SAT scores) to college admission officers in all future test reporting with this format.

It is ironic that if the College Board has altered the current test to reduce ethnic differences — a praiseworthy goal in itself — it may have inadvertently muted the possibility of showing really large R-SAT gain scores and hence less test bias for the minorities (by dropping all antonyms). Nevertheless, it is quite possible that large R-SAT scores might still be forthcoming with the current SAT format.

Concerns Regarding the Proposed New Format for the SAT[41]

How will the newly proposed changes in the SAT affect minority performance? Will something akin to an R-SAT score have any likely significance, given this newly proposed design?

The new SAT will probably consist of a quantitative section with several new items dealing with more advanced algebra content. Also, the few items dealing with so-called quantitative comparisons will be eliminated.[42] There will be a Critical Reading Test, essentially consisting of more reading items

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

with presumably shorter reading passages (i.e., each short passage will probably have just one or two items associated with it). There will be a Writing Section consisting of two parts: a written essay (perhaps the student will be allowed to select one topic from a list of possible topics) and a multiple-choice grammar test. Analogies will be eliminated (and, of course, there will be no antonyms, since these have already been dropped from even the most current SAT version).

Regarding the quantitative section, I predict that a separate R-SAT score for the math section will still be relevant; the more advanced algebra items (assuming these items are among the more difficult math items) are likely, given our earlier results, to favor the minority population, once the White and African American populations have been equated for their total SAT math scores. This assumes, of course, that the African American students will have been properly instructed in this more advanced subject matter in their high schools.

The issue concerning the elimination of the so-called quantitative comparison items should pose no problem for the math R-SAT score, since Kulick and Hu (1989) have already shown that the math items (with the quantitative comparison items deleted) yield the familiar DIF bias pattern, with African Americans and other minorities performing differentially better on the *hard* items.

Regarding the requirement for a written essay, data from a new study shows (Freedle, 2003) that African American, Hispanic American, and Asian American students perform significantly better on harder topics in comparison with their performance on easier topics.[43] That is, minority essay scores were more similar to those of White examinees for the hardest topics than for the easiest topics. Such a finding is similar to what Flaugher and Schraeder (1978) already reported for the old SAT. African Americans and Whites (without benefit of the DIF statistic) are more similar on *hard* SAT items than on *easy* items. The results for the written essay topics are part of a larger study that examines what factors influence topic difficulty for students taking the AP biology test over a ten-year period. Obviously, I must make the critical assumption here that the results found for the biology topics also will apply to essays written on more generally assigned topics; that is, topics that do not assume any specialized knowledge of a particular subject matter.

If I am correct in this assumption, then there should be interest in examining topic difficulty in order to select SAT essay topics in such a way that the several ethnic groups will be more similar in mean performance. There is less chance for ethnic bias to appear in the essay section of the SAT if hard essay topics are selected and if students are freely allowed to structure their essays with few restriction on what content to include.

Regarding the other changes that probably will be adopted for the new SAT — dropping the analogies, adding a grammar section, and including more reading items via shorter reading passages — I predict, based primarily

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

on Freedle and Kostin (1990), that these changes will not contribute in any significant way to the biased pattern I have described. Nevertheless, the new grammar section should still be empirically evaluated using the DIF statistic, as described earlier.[44]

In summary, regarding the newly proposed SAT, at least for the math sections and the essay requirement, it seems highly likely that something akin to an R-SAT score would be beneficial to minority students taking the new SAT. The bias that existed in earlier versions of the SAT is not necessarily decreased or eliminated in newer (and proposed) versions. Thus, R-SAT scores should still be calculated, even on the newer SAT.

Conclusion

Large-scale standardized testing may be at a crossroads in U.S. education (Gose & Selingo, 2001). The recent decision of several large state systems (California, Texas, and Florida) to consider dropping the SAT as a requirement for college application sounds a possible warning. Although some state and college officials say that race is not the basis for questioning the regular SAT as a requirement, other observers suggest that the unintended existence of ethnic bias in the SAT underlies their efforts to minimize the importance of the test in college admissions.[45] With this in mind, I offered some concrete suggestions for how to reinvigorate the traditional SAT using the R-SAT so as to remove a large part of its cultural and statistical bias; I call the new technique the revised SAT score, or R-SAT. I believe this new score reduces ethnic bias and therefore has the potential to increase dramatically the number of minority individuals who might qualify for admission into our nation's select colleges and universities. I further show that disadvantaged White examinees can also benefit from application of this revised score. The often dramatic increases in an examinee's score as revealed by his/her R-SAT score results from focusing attention on how the student performs on the *hard* half of the SAT test, be it the verbal or the quantitative parts of the test.

The present research and earlier studies (e.g., Flaugher & Schraeder, 1978) shows that African Americans, Hispanics, Asian Americans, and disadvantaged Whites perform differentially better on *hard* verbal and quantitative items. Indeed, except for reading items, this pattern of results seems to represent a widespread phenomenon for multiple-choice tests; I have found evidence for this bias pattern across a wide span of tests and age groups. The existence of this problem of bias (and its concomitant solution via the R-SAT score) has appeared in such diverse contexts as a vocabulary test for highschoolers (Raju et al., 1993), the paper-and-pencil version of the Graduate Record Examination (Freedle & Kostin, 1997), the SAT (both verbal and quantitative), several Advanced Placement tests, and even in some unpublished reanalysis I did of the responses of African American and White third graders to a multiple-choice test. In all these examples it appears likely that

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

cultural familiarity and semantic ambiguity play an important role in determining the relatively poor performance of minority groups on essentially the easiest test items.

A great deal of empirical work still needs to be completed. The necessary reliability and validity studies for the R-SAT score must be addressed. Although it is not optimal, at the very least one should study the R-SAT score using the current SAT format by restricting computation of the R-SAT score to just the *hard* analogies and *hard* sentence completions, because that is all there is left in the current test format from which to construct this promising new score.

Results show that the amount of verbal context plays an important role in the degree of cultural bias that occurs, at least in the verbal section: the greater the verbal context, the less the cultural bias. As noted earlier, Freedle and Kostin (1990, 1997) have hypothesized that cultural bias appears to be closely associated with differences in how common vocabulary words are perceived across minority and majority cultures so that *easy* verbal items (especially analogy and antonym items) show the most extensive cultural bias. *Hard* vocabulary appears to be less culturally sensitive (tends to have unique dictionary entries and to reflect concepts learned primarily from textbooks); numerous analyses show that minorities perform differentially better on these *hard* items. *Hard* items (not *easy*) therefore have a more curricular and achievement orientation.

Having said this, a general disclaimer is in order. For many years, before each new test is compiled, a strong and in many ways admirable effort has been made by ETS and the College Board to help identify individual items that produce very large ethnic differences. Psychometric techniques are applied, and these errant items are regularly discarded. In addition, special committees are convened to examine individual item content for obvious ethnic differences; such items are also discarded. Why then do the tests still show small, persistent, and highly patterned ethnic differences that cumulatively can have such an impact on estimated ability, such as those reported by Freedle and Kostin (1990, 1997)? Again, my answer is that one cannot fully erase the pervasive influence of cultural linguistic background when examinees are asked to process "common" words that occur on the test without ample context to help separate the various semantic senses of these common words. The various shades of meaning of common words are shaped by the cultures that make frequent use of these words. Cultural communities have variable needs, which are reflected in these vocabularies (Gumperz, 1982; Gumperz & Levinson, 1996; Hall & Freedle, 1975; Hall, Nagy, & Linn, 1984; Heath, 1983; Scarr, 1994; Schwanenflugel, Blount, & Lin, 1991).

Because obviously flawed items have been eliminated, the DIF magnitudes for individual items that remain in the test are not large — a difference usually at most of about four percentage points for an *easy* item (e.g., 80% correct for African Americans versus 84% for Whites). However, it should be

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

stressed that the accumulation of these small differences across a span of forty items can have a dramatic effect on an examinee's estimated true ability. I have argued that the solution is not to get rid of all such residual DIF or, more radically, to throw out the entire test; the solution is to recognize that this is a pervasive phenomena that can be easily remedied by reporting two scores, the usual SAT and the R-SAT. For outsiders to continue to blame the test industry as the cause of the ethnic score differences is unfair and simplistic. However, having said that, I certainly hope that the testing industry can be persuaded to carry out the crucial predictive validity studies suggested herein. Should the validity studies find this new scoring system meritorious, I would hope that the testing industry would implement the reporting of this additional Revised-SAT.

Indeed, it seems to me that the College Board and ETS are strongly obligated to investigate these several recommended revised scores in the next two years, before the new SAT is implemented, in order to conduct the relevant reliability and predictive validity studies of the R-SAT. The expense is truly minimal, the moral obligation maximal.

Notes

1. Most of the work in this article was completed while the author was a senior research psychologist at the Educational Testing Service, Princeton, New Jersey. He is now retired. None of the ideas expressed in this paper should be interpreted as representing any official position of either the Educational Testing Service or the College Board. They are strictly the viewpoints of the author.
2. To keep matters manageable, the first part of this report focuses on possible increases in the verbal SAT scores for primarily African American students. Of course it is true that having higher math scores (e.g., SAT-M scores) as well should be considered in studying possible increases in minority admissions into the select colleges. But for simplicity's sake, the SAT-M will be examined in a later section of this article.
3. The background questionnaire for the SAT inquires about a student's language strengths in the following way: "What language do you know best? A) English; B) English and another language about the same; C) Another language." I refer to students who select option "A" as those for whom English is their sole preferred language or for whom English is their best language (of course, it is possible that these students may either be monolingual or they may have limited competence in another language or languages). Those students who select "B" or "C" are clearly bilingual.
4. It is important to note that, unless otherwise indicated, all the DIF analyses reported below selected only examinees that listed English as their best language.
5. See Freedle and Kostin (1990), for the computational DIF formula.
6. "Differentially" implies that, while the scores of African American and White students may increase under the new estimation technique, the change in score is greater for African Americans than Whites.
7. Freedle and Kostin's (1988) work was replicated for African Americans by Kulick and Hu (1989). They also extended the earlier DIF findings to Hispanic Americans and Asian Americans and found identical bias patterns.
8. *Hard* and *easy* items are defined as follows: for a given test, if one rank orders the items by the percentage of all students that pass each item, then the items can be divided into

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

two halves, with 50 percent of the items yielding the largest percentage passing (the *easy* half) and the remainder representing the lowest percent passing (the *hard* half).

9. Grouping the analogies so that we get the average DIF value for all analogies from item difficulties ranging from 4.0 to 10.9, the single average value for these easier analogies equals –.027. If we group all the harder analogies from item difficulties ranging from 11.0 to 18.9 together, we get an average DIF value of +.012 (these grouped averaged values are also shown in Table 1).
10. Some readers may find it counterintuitive that, if Asian Americans score very well on SAT verbal and math items vis-à-vis the White majority, how can they be showing further positive gains on the *hard* verbal and *hard* math items? The quantitative details necessary for answering this question are presented later in this article when we examine the existence of the same bias pattern at *each and every* SAT score level. It then becomes evident that it is *not* the mean performance of each ethnic group that influences the emergence of a bias pattern, but rather whether there is a cultural mismatch between the White majority and any given ethnic minority.
11. I will use SAT-M to refer to quantitative items of the SAT.
12. Carlton and Harris (1992) reported additional replications of Kulick and Hu's SAT quantitative bias pattern (wherein minority groups differentially respond better to *hard* math items).
13. Some potentially relevant work in this regard is covered by Hall, Nagy, and Linn (1984), who contrasted African American and White children's language in a variety of settings (e.g., school versus home). Hall et al. compiled a separate dictionary of words used by each ethnic group; however, though this work is quite valuable in itself, the dictionary does not distinguish the different semantic senses with which each word is used within and across racial groups and so cannot directly help us solve the problem at hand.
14. Regarding lexical disambiguation, see Miller, Heise, and Lichten (1951) and Miller (1999).
15. The two halves together must in most cases mathematically balance out to yield the fixed score of 500 (for further comments on cognitive aspects of *hard* item performance see Freedle, Kostin, & Schwartz, 1987).
16. Briefly, for our purposes, computer-adaptive tests (CATs) differ from paper-and-pencil (P&P) tests in several respects. To evaluate a student's ability using a P&P test, the student is presented with a fixed set of items, and every correct item receives one point, regardless of the difficulty of the item (for convenience I'm ignoring omitted items and items not reached). In contrast, no fixed set of items is presented for a CAT. By following a complex algorithm, a student's ability is estimated sequentially, such that if a student responds correctly to an item of known difficulty, he/she is then presented with an even harder item from a large pool of available items. But, if the student misses the item, an easier item is next presented from the item pool, and so on. Gradually, the program is able to narrow down a score assignment by locating the hardest items within which a student can successfully respond, given some statistically desirable degree of confidence. For CAT, there is no simple relationship between the total number of items correctly responded to over the whole procedure and the final score assignment.
17. White test-takers are used as the normative group for two reason. First, they constitute the largest single ethnic group taking the SAT (e.g., in the 1980s they represented about 78% of the test-takers, while Blacks represented 9%). Second, ethnic comparison statistics (such as DIF) as computed at ETS typically use White test-takers as the reference group.
18. To carry out these analyses, I will use the formula that gives the most conservative results. Appendix A explains the relative merits of three computational formulas.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

19. For this same form, the highest-scoring White student scores at 600 or better on his/her R-SAT, starting with an original SAT score of 300.
20. The reader should note that had I used the recentered scale, all these revised scores would be even higher; for example, 600 would be 670 and so on. To keep matters simple, I am using the earlier unrecentered values since these data come from earlier test forms.
21. Just to be sure there are no unwelcome surprises, it probably would be useful (for a few test forms) to calculate two separate R-SAT scores for each examinee — one for the *easy* half and one for the *hard* half. Very high SAT performers necessarily must do well on the *easy* and *hard* halves; but if they do slightly better than expected on the *hard* half (given their total SAT score), they must have done ever so slightly worse somewhere on the *easy* items; some medium SAT performers (say, scores of 300–600) are also expected to show a mathematical symmetry — perhaps performing noticeably better on the *hard* (than expected from their total SAT score); hence, they should in turn be performing noticeably worse on the *easy*. The lowest SAT performers, however, can actually earn a total score that yields large negative numbers due primarily to overcorrection for guessing and/or a complex pattern of item omissions or items not reached (at the end of sections); such individuals collectively get a score of 200. It's possible that such individuals may occasionally show a distinct nonsymmetry in their R-SAT for the *easy* as opposed to the *hard* half of the test. In such cases college admission officers may want to see a separate breakdown of *easy*- versus *hard*-item performance, especially for those rare individuals who happen to score extremely well on just the *hard* half.
22. See the information posted on the College Board website at www.collegeboard.com.
23. See Table 1 for the general bias pattern.
24. Although the SAT may be losing its preeminence as a factor in college admissions decisions (see Gose & Selingo, 2001), I maintain the assumption here of its continued importance. I assert that correcting the SAT score (via the R-SAT score) will likely lead to more minority admissions into select colleges. I am fully aware that this may be true in a general sense — that having higher SAT (or R-SAT) scores certainly cannot hurt minority admissions, but that many other factors might figure in the decision process. This assumption is not without merit; Gose and Selingo suggest that the SAT's importance may persist because SAT scores figure prominently in annual college ranking guidebooks that presumably strongly influence college applicants. As such, select colleges may feel the necessity to require the SAT (and the R-SAT, hopefully), simply to help maintain their high rankings in these guidebooks.
25. Readers interested in seeing the details behind the calculation for estimating the number of African Americans whose R-SAT score equals or exceeds 600 at each of the SAT original score levels starting at 200 and moving up through 590 are referred to Appendix B.
26. These analyses of Forms OB023 and 4I were based on small samples taken from a large examinee pool. The actual frequency of African Americans who qualify for admissions would be correspondingly larger. Today, African Americans represent 11 percent of the SAT test-taking population. A total of 1,260,278 students took the SAT in 2000 (www.collegeboard.com), including about 139,000 African Americans. Of these, 0.72 percent, or 1,001 African Americans, were "qualified" for admission to a select college based on my hypothesis that a regular SAT score of at least 600 would qualify a student as a serious candidate. As shown in Table 4, when using combined R-SAT and SAT figures, 2.46 percent of African Americans score at or above 600. If we apply this percentage to the total figures for African American examinees cited above, it results in 3,419 African American students scoring at the 600 or above range, an increase of over 2,400 students potentially available for admission. These results hold considerable promise, I believe, for addressing the currently low admission rates of African Americans into se-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

lect colleges. Of course, it is also true that additional White students will be competing for admission to these same elect colleges due to the beneficial application of R-SAT scores for all White examinees. How admission officers will choose to balance these new possibilities remains to be seem.

27. Without another sample to replicate the pattern of these correlations, it is difficult to explain why the Asian Americans produced such a large significant correlation for the ninety sentence-completion items reported in Table 5. What is more important at this point is that the correlations all show the same persistent ethnic bias pattern — *easy* items are still differentially more difficult for each ethnic group, while *hard* items are differentially easier.
28. The matched Whites for all these cross-ethnic comparisons is always the same group of 5,147 Whites for whom English is their sole preferred language.
29. This latter finding is consistent with Freedle and Kostin's (1997) study that reported that abstractness (i.e., nonconcreteness) was one of the factors producing positive DIF for African Americans in responding to *hard* analogy items. Unfortunately, Carlton and Harris (1992) did not explicitly link their DIF findings to item difficulty per se, as Freedle and Kostin (1988, 1990, 1997) have done.
30. The Advanced Placement (AP) tests are generally taken by very able high school students. In 1997, for example, 68 percent of the AP test-takers had a grade point average of A– or better. In addition, 39 percent of all AP test-takers planned to get a doctoral degree (see www.collegeboard.com). Fifty-five percent were female, 71 percent White, 5 percent African American, 3 percent Hispanic or Latino, and 12 percent Asian, Asian American, or Pacific Islanders.
31. A DIF analysis of the several subject-area tests of the SAT II (i.e., the achievement tests) would have been desirable. However, these data are not available to the author.
32. In 2001, 252,504 students took one or more of the twenty-two listed subject area tests for the SAT II; of these, fully 225,724 took the English writing test. Of the total SAT II test-takers, 157,548 took the math level IC subject-area test. It is clear that these two subject-area tests are the ones most requested by various colleges to which these students have applied. It is my understanding that colleges that request these additional SAT II tests tend to be the more select colleges. We see, therefore, that approximately 20 percent of SAT I takers also take one or more of the SAT II tests. It is noteworthy that the mean verbal SAT I scores of students who take various SAT II tests are often considerably higher than the mean verbal scores for the student body who take the regular SAT I. For example, those taking the English writing test, the most popular subject-area test, earn a mean verbal SAT I of 599 SAT II, which is about 100 points above the total SAT I average; these same students' mean SAT I math score is also about 100 points above the mean of all SAT I takers math score. The students who take the SAT II tests are therefore a more select student population (see News & Information, www.collegeboard.com).
33. The reason the correlation is negative is that here, item difficulty is indexed by the so-called Equated Delta values for each item rather than percentage correct. Equated Delta is a statistic used by ETS in order to facilitate comparison across different test forms — Equated Delta adjusts the percentage corrects upward or downward for a given test form, depending upon the relative ability of the population taking the test at various times of the year; the results of these adjustments are placed on a new scale called the Equated Delta scale (see Kulick & Hu, 1989). In practice, the choice of which measure to base the correlation on (either the percentage correct or Equated Delta) does not seem to make that much difference; however, one can see that the use of the statistic called the Mantel-Haenszel for the two AP tests appears to magnify all the correlational effects — this statistic is much more sensitive to the "tails" of the distribution of proportions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

34. Traditionally, the raw test scores are corrected for "guessing" by use of the formula R – W/4, where R = number of correct items and W = number of wrong items. Suppose there are one hundred items and all items are randomly guessed; for a five-option item format, on average, one would expect twenty of these items to be correct by chance alone. Eighty items would then be wrong and this eighty is divided by four, equaling twenty. Applying the formula, the twenty correct minus the 80/4 fraction equals zero. By assumption, since the whole test was randomly responded to, the final score, corrected for chance, would be zero. I argue that this correction-for-guessing formula artificially penalizes minority test-takers. The key reason is that minorities get many *easy* items wrong because of a cultural ambiguity in interpreting many common vocabulary terms. These culturally induced wrong item responses get added to their other wrong answers (which are wrong, presumably, due to a true lack of information) and then, after dividing by four, are subtracted from their correct scores. It should be clear, then, that a bigger fraction is unfairly being subtracted from their correct scores due to the cultural mismatch that occurs primarily on the *easy* items. For this reason I recommend abandoning the correction for guessing formula before the final scaled scores are determined.
35. In 1987, White (non-Hispanic) students represented 78 percent of test-takers, while African Americans represented 9 percent; together they comprised 87 percent of the test-takers. By 1997, these percentages had changed so that White (non-Hispanic) students now represented 68 percent compared to 11 percent African Americans (a total of 79 percent) (see www.collegeboard.com).
36. Sigma is calculated by first determining the mean performance of each of two ethnic groups (here, the mean for African American performance and the mean for White performance). Then the two means are subtracted and the result is divided by the standard deviation for the combined group. This yields the final sigma value.
37. The reader may be puzzled as to why the African American mean for the *hard* items is not higher than the White mean; after all, many of our tables have shown that when equated for original SAT score, the African Americans almost invariably perform better on the *hard* items in comparison with their matched-ability White peers. The reason the African American *group* mean performance is lower than the White *group* mean is due to the number of examinees that occur at each scoring level; the number of African American examinees tends to be larger at the lower SAT levels while the number of White examinees tends to be larger for the higher SAT levels. These differential frequencies of occurrence result in an overall lower African American mean performance on the forty *hard* items in spite of their superior performance at each particular score level. The reader is referred to footnote 38, which describes a set of social variables that artificially lower the overall mean performance of African Americans. Once all these factors are controlled for, one might very well see the disappearance of any significant mean differences in ethnic SAT comparisons. Unfortunately, the author no longer has the raw data with which to evaluate this supposition.
38. It is important that the reader be informed about a few variables that unfairly affect mean test performance of disadvantaged groups. When population statistics are reported for contrasting, say, African American and White responses to verbal and/or quantitative items, there is little care given to the fact that these populations differ in a number of factors that contribute unfairly when mean performance levels are being considered. Some of the factors that have a negative impact on mean performance are family income level, amount of prior SAT practice, whether some language other than English is the primary means of expression, and the quality of education one has received. For African American examinees the presence of each of these factors artificially depresses their mean SAT performance with respect to White mean performance. For example, in my analyses of recent SAT forms, the biographical responses

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

show the following: 1) 9 percent of the African American population reports that a language other than English is their sole preferred language, while only 5 percent of the White population reports similar language preferences; 2) 36 percent of the African American examinees come from families earning $25,000 or less, while only 10 percent of African American applicants come from families earning $70,000 or more; on the other hand, only 12 percent of White examinees come from families earning $25,000 or less while almost 30 percent of the White examinees come from families earning $70,000 or more; 3) regarding prior practice on the SAT, Powers and Rock (1999) show that African Americans have less prior practice, which artificially lowers their mean performance; 4) African American examinees are generally acknowledged to receive a relatively poor education (Lemann, 1999). Each of these factors has a negative impact on mean SAT score level for African Americans, yet these many factors are seldom if ever partialled out or otherwise controlled for when official reports are issued that are intended to indicate overall mean ethnic performance of African Americans and other minorities. This is obviously unfair, inasmuch as the total impact of these several uncontrolled factors is, in my opinion, undoubtedly quite large. Studies typically will control just one of these factors at a time (e.g., Jensen, 1980) and will be comfortable demonstrating that large ethnic population differences in performance still remain; but virtually no study (including this one) has, to my knowledge, controlled for all these factors. College admissions officers are, of course, aware of the negative impact of these several factors on traditional SAT scores and has adopted various measures to try to ameliorate their effects on the admissions process. It is my expectation that the virtue of incorporating R-SAT scores into the decision process might provide a more objective way to bypass the negative impact of many of these uncontrolled factors.

39. For details, see www.collegeboard.com.

40. Generally speaking, the more items comprised in a test score, the more reliable that test is. Any such R-SAT score that is calculated using the old SAT data will have to be statistically increased via a "correction for attenuation" due to the fewer test items used in its calculation.

41. This section describes potential SAT changes discussed on the College Board website (www.collegeboard.org). These changes are expected to be implemented in 2004 and 2005.

42. These special items were not separately discussed above because they show the same DIF pattern as all the math items considered as a whole. They are mentioned here only because the College Board and ETS have specifically targeted them for elimination.

43. Freedle (2003) uncovered several significant factors that influence essay difficulty. The forty essay topics that were studied differed in a number of structural ways; for example (a) they varied in the amount of verbal background material that was presented in written form to students before the students wrote their essays; (b) the topics varied as to whether the background material was supplemented by graphs or tabular information or neither; and (c) the topics differed in how many specific subtasks were required of the student in order to successfully complete the essay assignment, and so on. Of eight factors examined, the three factors just mentioned contributed significantly to essay difficulty. The results showed that 1) the greater the amount of background information provided to students, the easier the essay; 2) if the background information was supplemented by tables or graphs, the essay topic was easier; and 3) the greater the number of specific subtasks requested to successfully complete the essay, the easier the essay. Put another way, *hard* essays had fewer guidelines in how to construct an acceptable essay — they provided less background information and provided fewer specific subtasks to cue the writer as to what a successful essay would consist of. Students writing the *hard* essays were more on their own in selecting relevant facts and in determining

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

how they would organize the information. As I have already indicated, minority students performed more similarly to White candidates on the hardest essays; that is, while Whites did in fact earn higher mean essay scores across the difficulty spectrum, the difference in mean performance between Whites and each minority group was smallest for the *hard* essays. If we assume that the results for the forty biology essays can indeed be generalized to essay writing across a variety of subject matters, then Freedle's (2003) results suggest that special caution will have to be exercised to ensure that the essay topics assigned to SAT examinees are sufficiently difficult so as to minimize mean score differences among all ethnic groups — that is, to minimize ethnic bias whenever possible.

44. Pragmatically speaking, the optimal way to design verbal, math, and essay items is to study how every item correlates with every grade obtained across a wide variety of college courses; if one selects the item types that yield the largest correlations, this necessarily increases the predictive validity of any test consisting of such items. For example, certain types of general math items are likely to correlate strongly with physics, chemistry, and math course grades; and certain types of general verbal (or essay) items might correlate strongly with particular literature or history courses. By increasing these special item types in the verbal and quantitative sections of any new SAT test, one can gradually improve the predictive validity of the SAT for each ethnic group. Such a selective process of gradual improvement in predictive power was actually used by Alfred Binet (1909) when he was developing the first standardized tests in France nearly a century ago. I believe there is wisdom in returning to his original method of test construction.
45. See Atkinson (in press).

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–106.

Atkinson, R. C. (in press). Achievement versus aptitude tests in college admissions: *Issues in Science and Technology.*

Berger, K. W. (1977). *The most common 100,000 words used in conversations.* Kent, OH: Herald.

Binet, A. (1909). *Les idées modernes sur les enfants.* Paris: Ernest Flammarion.

Breland, H., & Jenkins, L. (1997). *English word frequency statistics: Analysis of selected corpus of 14 million tokens.* New York: College Entrance Examination Board.

Carlton, S. T., & Harris, A. M. (1992). *Characteristics associated with Differential Item Functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons* (Research Report No. 92-64). Princeton, NJ: Educational Testing Service.

College Board. (1983). *10 SATs, Scholastic Aptitude Tests of the College Entrance Examination Board.* New York: Author.

College Board. (2001). Available online at www.collegeboard.com.

Cook-Gumperz, J. (1973). *Social control and socialization: A study of class differences in the language of maternal control.* London: Routledge & Kegan Paul.

Diaz-Guerrero, R., & Szalay, L. B. (1991). *Understanding Mexicans and Americans: Cultural perspectives in conflict.* New York: Plenum Press.

Dorans, N. (1982). *Technical review of SAT item fairness studies: 1975–1979.* Unpublished statistical report, No. ST-82-90. Princeton, NJ: Educational Testing Service.

Dorans, N., & Kulick, E. (1983). *Demonstrating unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the Standardization approach* (Research Report No. RR-83-9). Princeton, NJ: Educational Testing Service.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Flaugher, R., & Schraeder, W. (1978). *Eliminating differentially difficult items as an approach to test bias* (Research Bulletin No. RB-78-04). Princeton, NJ: Educational Testing Service.

Freedle, R. (1997). The relevance of multiple-choice reading test data in studying expository passage comprehension. *Discourse Processes, 23,* 399–440.

Freedle, R. (2003). *Structural factors that significantly influence essay difficulty: Implication for ethnic differences in standardized testing.* Unpublished manuscript.

Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. O. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 162–192). Norwood, NJ: Ablex.

Freedle, R., & Kostin, I. (1988). *Relationship between item characteristics and an index of Differential Item Functioning (DIF) for the four GRE verbal item types* (Research Report No. RR-88-29). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of Differential Item Functioning for Black and White examinees. *Journal of Educational Measurement, 27,* 329–343.

Freedle, R., & Kostin, I. (1991a). *The prediction of SAT reading comprehension item difficulty for expository prose passages* (Research Report No. RR-91-29). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1991b). *The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main idea, inference and explicit statement items* (Research Report No. RR-91-59). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1997). Predicting Black and White Differential Item Functioning in verbal analogy performance. *Intelligence, 24,* 417–444.

Freedle, R., Kostin, I., & Schwartz, L. (1987). *A comparison of strategies used by Black and White students in solving SAT verbal analogies using a thinking aloud method and a matched percentage-correct design* (Research Report No. RR-87-48). Princeton, NJ: Educational Testing Service.

Gose, B., & Selingo, J. (2001, October 26). The SAT's greatest test. *Chronicle of Higher Education,* p. A10.

Gould, S. J. (1995). Mismeasure by any measure. In R. Jacoby & N. Glauberman (Eds.), *The bell curve debate* (pp. 3–13). New York: Random House/Times Books.

Gumperz, J. J. (1982). *Discourse strategies.* New York: Cambridge University Press.

Gumperz, J. J., & Levinson, S. C. (Eds.). (1996). *Rethinking linguistic relativity.* New York: Cambridge University Press.

Hall, W. S., & Freedle, R. (1975). *Culture and language: The Black American experience.* Washington, DC: Hemisphere/Wiley.

Hall, W. S., Nagy, W., & Linn, R. (1984). Oral language data and sociolinguistic variation. In W. S. Hall, W. Nagy, & R. Linn (Eds.), *Spoken words: Effects of situation and social group on oral word usage and frequency* (pp. 454–465). Hillsdale, NJ: Erlbaum.

Heath, S. B. (1983). *Ways with words.* Cambridge, Eng.: Cambridge University Press.

Herrnstein, R. J., & Murray, R. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

Hill, C., & Larsen, E. (2000). Children and reading tests. In *Advances in discourse processes* (series vol. 65). Stamford, CT: Ablex/Greenwood.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. Baun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1–123.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty* (College Board Report No. CB-89-5). New York: College Entrance Examination Board.

Labov, W. (1975). Academic ignorance and Black intelligence. In M. Maehr & W. Stallings, (Eds.), *Culture, child, and school: Sociocultural influences on learning* (pp. 63–81). Monterey, CA: Brooks/Cole.

Lemann, N. (1999). *The big test.* New York: Farrar, Straus & Giroux.

Miller, G. A. (1999). On knowing a word. *Annual Review of Psychology, 50,* 1–19.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a functioning of the context of the text material. *Journal of Experimental Psychology, 41,* 329–335.

Mosteller, F., Rouke, R. E. K., & Thomas, G. B., Jr. (1961). *Probability and statistics.* Reading, MA: Addison-Wesley.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36,* 93–118.

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test and the Mantel-Haenszel techniques for assessing differential item functioning. *Educational and Psychological Measurement, 53,* 301–314.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Students group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report No. 93-1). New York: College Entrance Examination Board.

Scarr, S. (1994). Culture-fair and culture-free tests. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of paper-and-pencil and computer adaptive test scores on the GRE general test* (Research Report No. 98-38). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1991). *Differential speededness and item omit patterns on the SAT* (Research Report No. RR-91-50). Princeton, NJ: Educational Testing Service.

Schwanenflugel, P., Blount, B., & Lin, P. (1991). Cross-cultural aspects of word meanings. In P. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 71–90). Hillsdale, NJ: Erlbaum.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14,* 181–196.

Wilson, K. (1988). A study of the long-term stability of GRE general test scores. *Research in Higher Education, 29,* 3–40.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Appendix A:
The Effect of Three Formulas for Calculating DIF Scores

There is a technical issue that concerns the choice of three possible formulas for calculating how well a student has performed on a multiple-choice test (be it just for *hard* items or for the full set of test items). The issue arises of which formula is most appropriate for DIF studies because a few of the terms that occur in the several formulas can show a different rate of occurrence when comparing across the various minority groups. The question naturally arises as to whether these different rates seriously affect the conclusions one might draw regarding DIF comparisons.

Each of the three formulas involves calculating a percentage correct score using the number of correct responses (designated "R" for number right) in the numerator, but differing in what is placed in the denominator. There are three possible selections for a denominator: R+W, R+W+O, or R+W+O+NR, where R = number of correct answers, W = number of wrong answers, O = number of omitted items (these typically are blank items that occur among the correct and incorrect answers), and NR = number of items not reached (as might occur at the end of a test section). Let us focus on just the forty hardest items in calculating a score. For example, if R = 20, W = 10, O = 5, and NR = 5, the three formulas would yield the following 3 percent correct responses for the very same individual: 20/(20+10) = 20/30 = 67 percent for the R+W formula. For the R+W+O formula, the percentage of correct responses would be 20/(20+10+5) = 20/35 = 57 percent. And for the R+W+O+NR formula, the percentage of correct responses would be 20/(20+10+5+5) = 20/40 = 50 percent. So there are three different ways to think about how well a particular individual has performed on the hardest forty items. How should one go about selecting a formula?

The earlier work of Freedle and Kostin (1988, 1990) used the R+W+O formula because that was the one traditionally used by ETS in many of its DIF analyses to identify highly divergent items during pretesting. However, Kulick and Hu (1989) suggest that, since Whites actually have a higher item omission rate than African Americans do on the SAT, this fact would artificially depress the performance of Whites on the *hard* items relative to African Americans. In fact, they suggest that if Freedle and Kostin had used just the R+W formula they would not have found any systematic bias effects. Because of this contention of Kulick and Hu, I have chosen to present the data in Table 2 (see main text) using just the R+W formula; yet this table clearly reveals the superiority of African American examinees on the hardest forty items at every calculated score level shown. Hence, Kulick and Hu, based on this evidence alone, are mistaken that the R+W formula will fail to show any systematic ethnic differences. Later, Schmitt et al. (1991) also showed that the Freedle and Kostin (1988) bias effect will persist no matter which of the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

three formulas is used. Therefore, in several sections of the main article I have used the more conservative score R+W+O+NR when I have determined that it is preferable to present conservative estimates, especially when it comes to estimating the percentage of African Americans who might qualify for admission to select colleges.

To further settle this issue, I present several new findings. Immediately below I present the average gain scores for three SAT tests using each of the three formulas. The mean African American gain scores for Form OB023 were 38.76, 30.08, and 30.09 for the three formulas R+W+O, R+W, and R+W+O+NR, respectively. For Form 4I, the comparable African American mean gain values were 30.26, 20.75, and 24.06. And for Form 4W the comparable values were 21.93, 17.88, and 16.09. (The number of African Americans for Forms OB023, 4I, and 4W were 7468, 9559, and 6225, respectively. For White students the corresponding numbers for these three test forms were 40102, 31487, and 37639.)

These results clearly show that large average positive gain scores occur for African American examinees for all three test forms and for each of the three formulas. Therefore, I conclude that omission rates, while undoubtedly differentially represented in the African American and White examinee populations, are not themselves the basic source of difference between the populations.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Appendix B:
Calculating the Number of African American Examinees Whose R-SAT Equals or Exceeds 600

Table B-1 lists the mean percentage (out of 40 items) that the White examinees and the African American examinees got right for each SAT score level (actually, for convenience, the score levels start at 470). Then it lists the standard deviation (S.D.) for all the White students at a given score level, and the S.D. for all the African American students at the same SAT level. Then it lists the best-scoring White and African American student (the student who got the highest percentage of the forty items correct for a given ethnic group). And, finally, at a given score level, it lists how many White and African American students obtained that original SAT score. I use the R+W+O+NR formula here because it gives the most conservative estimates.

Now I will show how to estimate the number of minority students who get a revised R-SAT score of 600 or better using these entries. First of all, note what the target is: we know from Table 6 that Whites who got an original SAT score of 600 also got 52.9 percent of the forty *hard* items correct. We want to determine how many African American students who originally got an SAT score of less than 600 are found to have actually achieved as good a performance on the *hard* half of the test as the White students did; ultimately, by applying our method to every score level we want to find how many African Americans at each score level between 200 and 590 achieved 52.9 percent (or more) of the *hard* items correct.

As a convenience and merely for the purposes of exposition, in what follows below I assume a normal distribution of scores around the mean percentage correct for African American and White students (remember, the final operational procedure for R-score assignment is outlined on pages 9–11). To begin with, note that at a score level of 470 there are sixty-four African American students represented. Half of these students (64/2 = 32) exceed the African American mean of 28.1 percent in their performance on the forty *hard* items. We already know that at least one student (the maximum scorer with 74.6%) exceeds the target of 52.9 percent, so we are certain that this maximum performing student will definitely get an R-SAT score of 600 or better (actually, this student's R-SAT score is 690). But how many other African American students who scored above the mean of 28.1 percent reached the target value or exceeded it? To find out, we use the S.D. of the African American students at 470, which is 16.1. We want to find a "z score" in standard deviation units that equals the target of 52.9 percent. The following calculation will estimate this for us.

$$28.1 + 16.1z = 52.9$$
$$z = 1.54$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE B–1 *Mean, Standard Deviation, and Maximum for Students' Percentage Performance on 40 Hardest Verbal Items*

*Original SAT Score**	*Type of Statistic*	*Whites*	*African Americans*
470	Mean	27.7%	28.1%
	S.D.	15.1	16.1
	Maximum Examinee	74.1%	74.6%
	Freq. Examinees	949	63
480	Mean	28.1%	29.8%
	S.D.	15.6	17.3
	Maximum Examinee	75.0%	81.7%
	Freq. Examinees	1630	93
490	Mean	30.3%	31.6%
	S.D.	16.0	18.7
	Maximum Examinee	78.9%	83.0%
	Freq. Examinees	885	47
. . .			
600	Mean	52.9%	(Not applicable)

* This table uses SAT Form OB023 and the formula R+W+O+NR for scoring (see Appendix A).

The first entry (28.1) is the average African American performance on the *hard* items for all African Americans who originally got an SAT score of 470. The next entry 16.1 is the S.D. for all the African Americans at SAT = 470. The "z" value, when solved for, gives us the number of S.D. "units" that are needed to help us reach the target performance of 52.9 (which is the White performance on the 40 hardest items whose original SAT score is 600).

A z score of 1.54 is the solution. Mosteller, Rourke, and Thomas' (1961, Table 3, p. 368) table tells us that a z score of 1.54 includes 43.82 percent (out of a maximum of 50%) who are above the mean. We subtract 43.82 from 50.00 = 6.18 percent. We then multiply 6.18 times 64 (the total number of African Americans at original SAT level of 470) to find out how many African American students get a score of 600 or higher. The solution is 4.0 (since 6.18% x 64 = 4.0). So, we estimate that a total of 4.0 African American examinees who originally scored 470 on the SAT actually have earned a score of 600 or higher for their R-SAT score (this estimate includes the maximum African American student with an R-SAT score of 690 mentioned above).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

A similar calculation is made for the White students; we want to see how many White students at each score level starting with 200 up through 590 have earned an R-SAT score of 600 or better on just the forty hardest verbal items. The critical level is again 52.9 percent. Any White student who equals or exceeds that level will get an R-SAT score of 600 or higher. (To show exactly what score level is achieved by each examinee would require a different computer program than the one currently available to me; that is why I have employed the current shortcut procedure for estimating R-SAT.)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.